Small Object Detection in Unmanned Aerial Vehicle Images Using Feature Fusion and Scaling-Based Single Shot Detector With Spatial Context Analysis

Xi Liang, *Member, IEEE*, Jing Zhang, *Member, IEEE*, Li Zhuo, Yuzhao Li, and Qi Tian, *Fellow, IEEE*

***Abstract*— Objects in unmanned aerial vehicle (UAV) images are generally small due to the high-photography altitude. Although many efforts have been made in object detection, how to accurately and quickly detect small objects is still one of the remaining open challenges. In this paper, we propose a feature fusion and scaling-based single shot detector (FS-SSD) for small object detection in the UAV images. The FS-SSD is an enhancement based on FSSD, a variety of the original single shot multibox detector (SSD). We add an extra scaling branch of the deconvolution module with an average pooling operation to form a feature pyramid. The original feature fusion branch is adjusted to be better suited to the small object detection task. The two feature pyramids generated by the deconvolution module and feature fusion module are utilized to make predictions together. In addition to the deep features learned by the FS-SSD, to further improve the detection accuracy, spatial context analysis is proposed to incorporate the object spatial relationships into object redetection. The interclass and intraclass distances between different object instances are computed as a spatial context, which proves effective for multiclass small object detection. Six experiments are conducted on the PASCAL VOC dataset and the two UAV image datasets. The experimental results demonstrate that the proposed method can achieve a comparable detection speed but an accuracy superior to those of the six state-of-the-art methods.**

***Index Terms*— Unmanned aerial vehicle (UAV) image, small object detection, feature fusion, feature scaling, single shot detector, spatial context analysis.**

Manuscript received August 29, 2018; revised January 20, 2019 and February 17, 2019; accepted March 8, 2019. Date of publication March 20, 2019; date of current version June 4, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61531006, Grant 61602018, and Grant 61701011, and in part by the Beijing Municipal Natural Science Foundation Cooperation Beijing Education Committee under Grant KZ 201810005002 and Grant KZ 201910005007. This paper was recommended by Associate Editor G.-J. Qi. *(Corresponding author: Jing Zhang.)*

X. Liang, J. Zhang, and Y. Li are with the Beijing Key Laboratory of Computational Intelligence and Intelligent System, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China (e-mail: liangxi627@emails.bjut.edu.cn; zhj@bjut.edu.cn; liyuzhao@emails.bjut.edu.cn).

L. Zhuo is with the Beijing Key Laboratory of Computational Intelligence and Intelligent System, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China, and also with the Collaborative Innovation Center of Electric Vehicles in Beijing, Beijing 100124, China (e-mail: zhuoli@bjut.edu.cn).

Q. Tian is with the Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX 78249-1604 USA, and also with the Noah's Ark Laboratory, Huawei Technologies, Shenzhen 518129, China (e-mail: qi.tian@utsa.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2019.2905881

I. Introduction

With the advantages of high mobility, fast deployment and large scope for surveillance, unmanned aerial vehicle (UAV) photography has been applied in many areas such as security and surveillance [1], search and rescue [2], sports analysis [3], etc., and UAV photography has become a good supplement to traditional remote sensing [4]. Due to the advantages of low cost, small volume, flexibility and convenience, UAVs as the primary low-altitude mobile platform can obtain information from airborne remote sensing equipment. After a series of computer processing steps, including geometric correction, image enhancement, image mosaic, etc., UAV images are acquired according to a certain accuracy [5]. In addition, the characteristic of low altitude makes UAVs not easily affected by clouds or other interference factors. Thus, UAV photography is a powerful supplement for satellite remote sensing and airborne remote sensing.

As a core problem in computer vision, object detection in UAV images is under extensive research in both academia and real-world applications, such as transportation surveillance, smart city, sports analysis, etc [6]. With the development of aerial photography and remote sensing technology, UAV images show the characteristics of massive data volume, multiperspectives and centimeter-level resolution. The objects in UAV images are usually small or tiny with ambiguous boundaries coupled with complex backgrounds and changing illumination conditions. Consequently, in the face of such complex and massive UAV images, how to quickly and accurately detect small objects in UAV images has both theoretical significance and practical application value [7].

Most of the traditional object detection methods in UAV images are based on the sliding-window paradigm, using hand-crafted features such as Histogram of oriented Gradient (HoG) features [8], Scale-Invariant Feature Transform (SIFT) features [7] and Haar-like features [9], which are time-consuming and laborious to achieve the robustness of feature representation. These low-level automation methods usually cannot meet the requirements of real-time object detection. More recently, object detection methods have made great progress with the resurgence of variety deep networks, including deep belief networks (DBNs) [10], convolutional neural networks (CNNs) [11], generative adversarial networks (GANs) [12], deep transfer networks [13], [14], etc. CNNs are a category of neural networks that have proven very effective

in multiple areas in computer vision. In the field of object detection, CNN-based modern detectors can generally be divided into the following two categories: two-stage detectors such as R-CNNs [15] and its many variants (Fast R-CNN [16], Faster R-CNN [17], R-FCN [18], etc.), and one-stage detectors (YOLO [19], SSD [20], etc.). Two-stage detectors divide the object detection task into the following two subtasks: identifying image regions that may contain objects, and then classifying each region individually. Thus, two-stage detectors are accurate but relatively slow, while the one-stage detectors aim to train a CNN model to map image pixels to coordinates of bounding boxes directly. Compared with two-stage detectors, one-stage detectors are more efficient in terms of both speed and memory but sacrifice the accuracy to some extent, especially in small object detection. Existing object detection methods are trade-offs between speed and accuracy [21]. In general, deep learning-based methods are computationally expensive in time and network volume but have relatively low accuracy for small object detection. In addition, the psychological evidence has suggested that context is vital for humans to recognize objects [22]. The empirical studies in the field of computer vision also proved that the performance of algorithms can be improved by proper modeling the spatial context whether by traditional method [23] or deep learning-based method [24]. Thus, the spatial context should have positive impact on small object detection in UAV images.

To tackle the problem of small object detection in UAV images, we propose a feature fusion and scaling-based single shot detector (FS-SSD) with spatial context analysis in this paper. By finding a better design of the feature fusion module and adding the deconvolution module with the average pooling layer, two feature pyramids are generated in the proposed FS-SSD to detect the small object. To make use of the interplay between different objects, spatial context analysis for object redetection is performed by computing the interclass and intraclass distances between object instances, which can further improve the detection accuracy for multiclass object detection.

The main contributions of this paper can be summarized as follows:

(1). By leveraging the speed and accuracy, a feature fusion and scaling-based single shot detector (FS-SSD) is proposed. We adjust the feature fusion module by adding an extra branch of the deconvolution module together with average pooling on the basis of FSSD. The incorporation of average pooling can help to prevent network overfitting to some extent by reducing the total number of parameters as well as by providing the background information of the image. Compared with the common bilinear interpolation operation for upsampling, the additional scaling branch with deconvolution module introduces the nonlinearity to the network to enhance the network representation ability. Two feature pyramids generated from the feature fusion and deconvolution module are used to make predictions for small objects.

(2). Since spatial relationship is ignored in most object detectors, a spatial context analysis method for object redetection is proposed in this paper. By taking the interplay of multiclass objects within a certain distance into consideration, interclass and intraclass distances between different object instances are computed as the spatial context to reverify the confidence of the existence of certain object instances. This redetection method helps to deal with multiclass small object detection by making full use of the spatial relationship, which can effectively improve the detection accuracy.

The rest of this paper is organized as follows. Sections II-IV describe the details of our method, including feature fusion and scaling-based SSD (FS-SSD) and spatial context analysis for object redetection. Section V presents the experimental results obtained from experiments conducted on the subset of the Stanford Drone Dataset (SDD) and CARPK dataset to validate the effectiveness of our proposed method. Section VI discusses the experimental results, our findings and future work directions, respectively. Finally, Section VII concludes this paper.

II. Overview of the Proposed Method

To improve the detection accuracy of the original SSD network for small objects, in this research, we make an improvement in the design of the deep learning model and develop an optimization method for the detection results. As depicted in Fig. 1, the proposed methods include the following two phases: (1) at the training phase, the feature fusion and scaling-based single shot detector is pretrained on PASCAL VOC [25] and COCO datasets [26] and fine-tuned on the experimental UAV datasets; and (2) at the detection phase, small object detection is achieved by the proposed FS-SSD model. To further improve the detection accuracy, spatial relationships between multiclass objects are utilized in the spatial context analysis for object redetection.

III. Feature Fusion and Scaling-Based Single Shot Detector (FS-SSD)

In this work, we developed certain design principles not only for compact architecture but also for accuracy of detection, as accuracy is of prime importance for us. We drew inspiration from NIN [27], DSSD [28], and the recent work of DetNet [29], designing an architecture that is effective for small object detection.

A. Multi-Scale Feature Maps for Prediction

Scale variation, a critical challenge in single shot multibox detectors, has a major effect on detection accuracy. As shown in Fig. 2, many methods have been proposed to solve the multiscale problem. In Fig. 2(a), the previous work based on hand-crafted features utilized multiscale images as input to generate different scale feature maps, which is quite inefficient. Fig. 2(b) uses the topmost feature map to create anchors with different scales by adopting some two-stage detectors, such as Faster R-CNN [17] and R-FCN [18]. However, the fixed receptive field of a single layer has difficulty in detecting multiscale objects. Fig. 2(c), adopted by the original SSD [20],

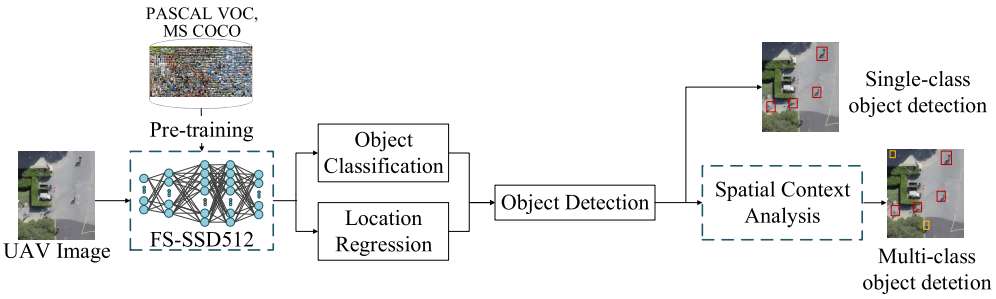


Fig. 1. Architecture of the proposed small object detection method in UAV images.

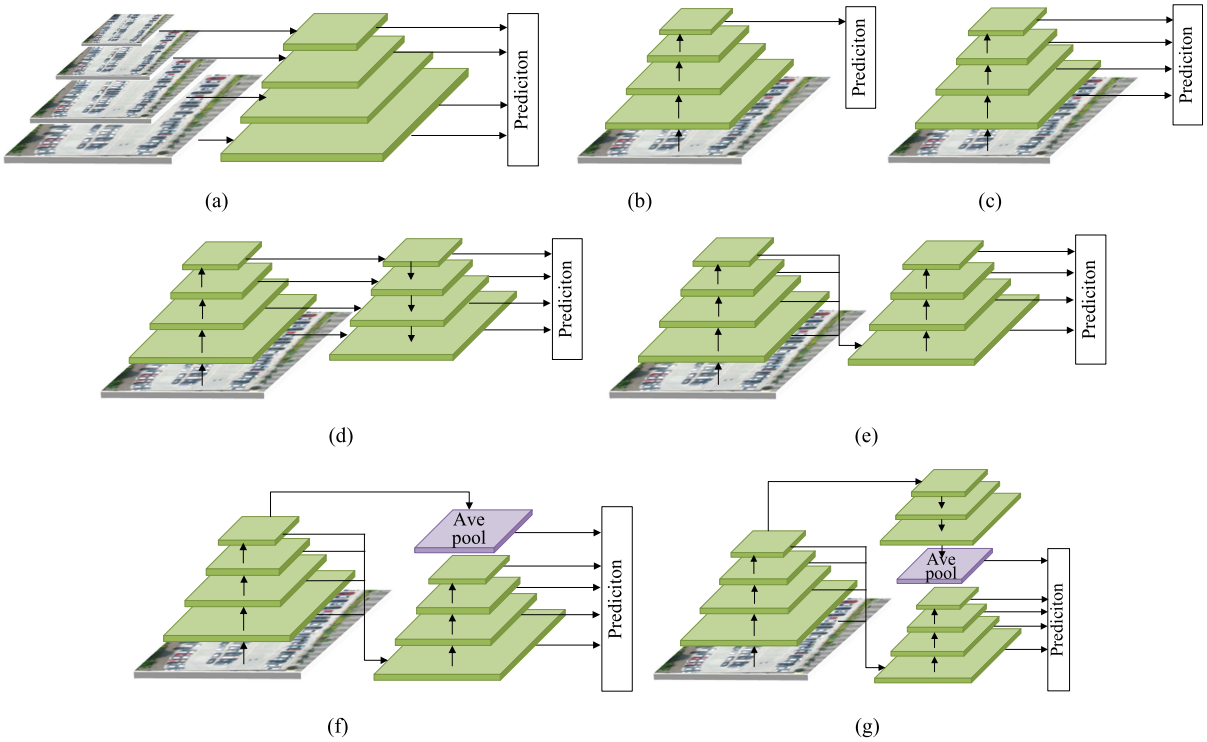


Fig. 2. (a) Features are computed from the image pyramids independently, which is computationally expensive. (b) The single-scale feature is utilized to make predictions, which are used in some two stage detectors such as Faster R-CNN and R-FCN. (c) Use of the feature pyramid generated from a single CNN. The conventional SSD is one of the examples. (d) Features are fused from top to bottom, layer by layer, which is adopted by FPN. (e) Features from different layers with different scales are concatenated together and used to generate pyramid features in FSSD. (f, g) Our feature fusion and scaling method. (f) A variant to (e) by adding an average pooling layer and adjusting the feature fusion module. (g) Feature pyramid after fused features as well as features amplified by the deconvolution module are used to make predictions together.

uses the feature pyramid from bottom to top to make predictions. In Fig. 2(d), FPN [30] developed a bottom-up and top-down architecture with lateral connections to get semantically stronger features, but it is not efficient enough to fuse features layer by layer. FSSD [31] combined SSD with FPN shown in Fig. 2(e), in which the feature pyramids are generated to make prediction after features are fused from bottom to top. However, the method of feature fusion in FSSD will cause the loss of low-level information. To deal with the problem of scale variations in a better way, we combine the advantages of the existing methods, proposing the feature fusion and scaling-based SSD (FS-SSD). Fig. 2(f) is our feature fusion-based SSD with average pooling. We add an average pooling layer to the final prediction layer at the end of network, as well as adjusting the feature fusion module. In Fig. 2(g), we further scale the features with the deconvolution module to increase the resolution of the feature map before average pooling. The feature fusion module and the deconvolution module make up the scaling module together to generate two feature pyramids to make predictions. Detailed information of the proposed network will be discussed in the following parts.

B. Feature Fusion Module and Average Pooling

The original SSD is built on the VGG network, which is truncated with some convolutional layers. As shown in Fig. 3(a), SSD adds several extra convolutional layers on top of the backbone network. Each of the added layers and Conv4_3 layer from the VGG network are used to make predictions. Non-maximum suppression (NMS) is used as the postprocess to get final detection results. More details can be found in [20]. Although SSD performs well in terms of

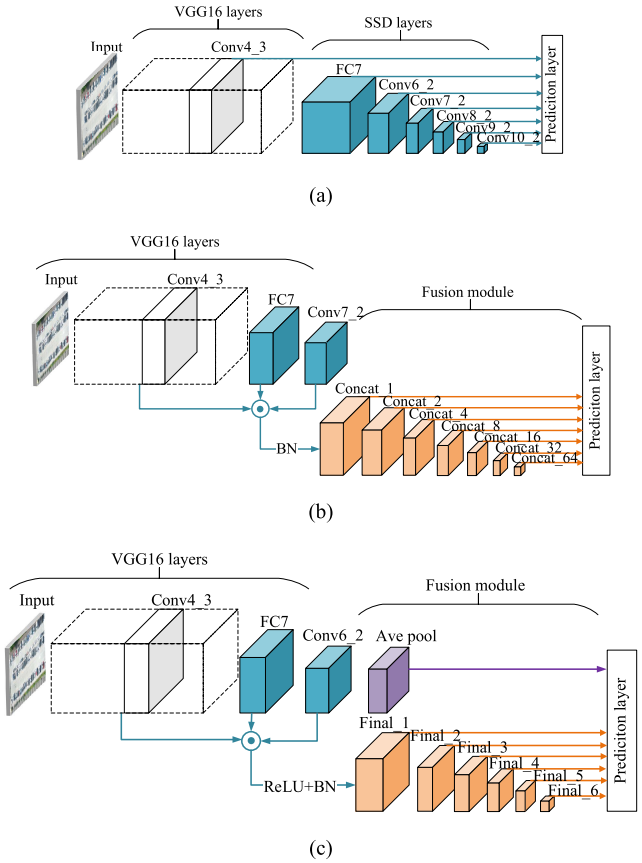


Fig. 3. The network architecture. (a) Original SSD network; (b) FSSD network; (c) Our network with feature fusion module and average pooling layer.

both detection speed and accuracy on most of the natural images, it is not suitable for small object detection. Since pyramidal features are used for prediction independently in SSD, the context information between different feature layers is not considered, which can greatly contribute to small object detection. Thus, in Fig. 3(b), FSSD is proposed to combine feature maps from different levels and generate a feature pyramid to make predictions. The experimental results show that the feature fusion module in FSSD can work better than FPN. However, the convolutional kernel sizes, strides, and paddings in FSSD bring some information loss from feature maps, which will influence the performance on small object detection. The final prediction layers also need careful redesign for our tasks. Therefore, in Fig. 3(c), we try to acquire the context information before feature fusion by adding an average pooling layer and to find the best settings to fully utilize the pyramidal features.

1) Average Pooling: Pooling, a common operation in a CNN-based backbone network, can achieve invariance in image transformation, making representations more compact and producing a higher receptive field that is beneficial to visual classification [32]. However, the spatial resolution is compromised and fails to localize and recognize the small objects accurately. The mainstream pooling operations include max pooling and average pooling. Max pooling chooses the discrete maximum on the pixel grid. When the maximum value of the features happens to be in the middle of all pixels, which often occurs during the training process, the selected maximum is not actually the true maximum, which will decrease the object classification accuracy. To minimize this situation and to retain more information from the feature neighborhood, we adopt average pooling rather than max pooling.

2) Feature Fusion Module: As mentioned above, there are many methods to utilize hierarchical pyramidal features fully by means of feature fusion. We follow the idea adopted by FSSD in Fig. 2(e) to save computational cost. Features are fused from different levels once in an appropriate way, and then feature pyramids are generated from the fused features. There are two main methods to merge different feature maps together, which are concatenation and element-wise summation. Concatenation requires only equal size in channel number between two feature maps, while element-wise summation requires feature maps to have the same size. We have compared the detection performance when using element-wise summation and concatenation to merge different feature maps together and finally chose concatenation to combine features due to its flexibility and superior performance. The detailed experimental results can be found in Section VB.

Feature selection is also an important factor in the performance of object detection. As shown in Table I, in the original SSD512, Conv4_3, FC_7 of the VGG16 and newly added layers Conv6_2, Conv7_2, Conv8_2, Conv9_2, Conv10_2 are chosen to make predictions. The feature map selection in FSSD follows the settings in the original SSD512. According to the analysis in [31], a feature map with spatial size smaller than 10 px $\times$ 10 px has little information to merge. Therefore, in the proposed FS-SSD512, Conv4_3, FC_7 and Conv6_2, whose feature maps are larger than 10 px $\times$ 10 px, are concatenated together to form the first final prediction layer, Final_1. Five selected feature maps, Final_2 to Final_6, are convolutional layers with high-level semantic information, which have the same resolution that is in FSSD512. We replace the 1 px $\times$ 1 px feature map with the average pooling layer feature whose spatial size is more suitable for small object detection.

C. Deconvolution Module

Based on the network in Fig. 3(b), we add a deconvolution module before average pooling to increase the resolution of feature maps successively, as shown in Fig. 4, which proves to be effective on both datasets in the following experiments. The deconvolution module consists of three 2 $\times$ 2 deconvolutional layers with stride 2 and one 3 $\times$ 3 convolutional layer, in which each layer is activated by Rectified Linear Units(ReLU), followed by a batch normalization operation. Compared with the traditional bilinear upsampling, deconvolutional layers contribute to the network feature representative ability by introducing the nonlinearity. The convolutional layer plays a role as the buffer, preventing the severe influence of the gradient from the whole network, maintaining the network stability [21]. The batch normalization layer adjusts the distribution of the input value of arbitrary neurons of every layer in the network to standard normal distribution with a

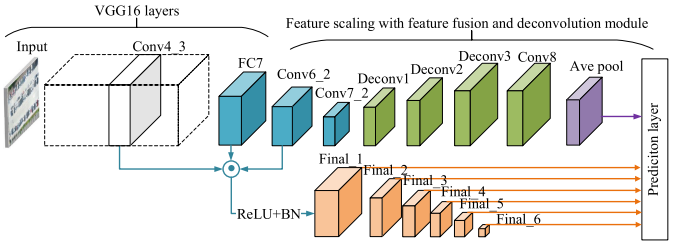


Fig. 4. Our proposed FS-SSD network.

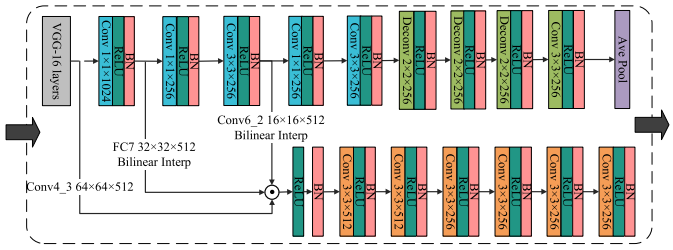


Fig. 5. Detailed structure of the additional module.

mean of 0 and variance of 1. It can make the activation value fall into the sensitive area of the nonlinear function to the input, so that the small changes in the input will lead to a larger change in the loss function and make the gradient larger. In this way, the problem of gradient disappearance will be relieved, and training time will be shortened at the same time [33]. The deconvolution module scales up the feature maps and the extra branch after the feature fusion module scales down the feature maps to generate two feature pyramids. The deeper architecture of the VGG16 backbone network together with the scaling module help to achieve higher accuracy while the narrow structure with two branches can better control the complexity of the network. Since we need to use dense feature layers to predict smaller objects, the last few feature layers after the feature fusion module as well as Conv4_3 layer are selected as the final prediction layer.

Fig. 5 illustrates the detailed design of the additional module in the proposed FS-SSD. To keep the efficiency of the proposed model, inspired by the design concept in [34], we employ a low complexity bottleneck layer with 1×1 convolution operation before each additional convolutional layer. The 1×1 convolutions are used to reduce the feature dimensions, to reduce the number of operations in each layer, saving computing costs and speeding up the reasoning process of the model. They can also provide the nonlinearity to the network. Each convolutional and deconvolutional layer is followed by a ReLU activation and batch normalization operation. By integrating the above improvements, our FS-SSD not only maintains high resolution feature maps but also keeps large receptive fields, both of which are important for small object detection.

IV. Spatial Context Analysis for Object Redetection

Objects in UAV images are usually small, which make it difficult for deep neural networks to capture the detailed semantic information, particularly if multiclass objects exist in UAV images. Although the proposed FS-SSD can utilize the context information by two feature pyramidal structure to some extent, it is not accurate enough to classify an object only relying on its deep features, especially when the object is too small or the features of the object are not robust enough. In order to make an improvement of the less reliable detection results, inspired by Leng and Liu [35], we propose the spatial context analysis to incorporate multiclass object spatial relationships for object redetection. Object detection is largely influenced by environmental factors, especially the surrounding objects. In the actual road scene, objects belonging to the same or similar classes tend to be closer to each other than objects with different classes. For example, under most circumstances, the pedestrians are at the side of the road whereas cars are at the center of the road that is relatively far from the pedestrians. Thus, we hope to make use of the interplay between different object instances within a certain distance to improve object detection accuracy under the certain confidence threshold.

The overall idea of spatial context analysis is to utilize the detected objects with high category confidence score to redetect the less reliable objects. For a less reliable object a, if there exist reliable objects within certain distance, the existence of a at current location proved more reliable, and the category confidence score of a will be increased. On the contrary, if there is no reliable object around a, the existence of a becomes less likely, and the category confidence score of a will be decreased.

By referencing the double threshold settings in Faster R-CNN [17], in the detection results of proposed FS-SSD model, we divide the detected objects into three groups: reliable objects, less reliable objects and unreliable objects. We assume an object is reliable when its category confidence score is larger than 0.6. Accordingly, when the category confidence score is smaller than 0.4, the object is regarded as unreliable. The redetection procedure aims mainly to improve classification accuracy of the less reliable objects whose category confidence score is between 0.4 and 0.6. Spatial context analysis is conducted using the following steps:

Step 1: For a less-reliable object set $\{a_i\}$, if reliable objects sets $\{b_j\}$ and $\{c_z\}$ exist whose distance to the object instances in $\{a_i\}$ is within d pixels, b_j has the same class with a_i, and c_z belonging to another class, then turn to *Step 2*. Otherwise, turn to *Step 3*.

Step 2: Since there exist reliable objects b_j and c_z around a_i, the possibility of the existence of a_i increased, and its final category confidence score C' will be raised according to the weighted distance between a_i and b_j, a_i and c_z, which is computed as (1):

$$C'(a_i) = C(a_i) + \lambda \times D(a_i, b_j) + (0.4 - \lambda) \times D(a_i, c_z) \tag{1}$$

where C is the confidence score of a_i after FS-SSD detection and D is the normalized distance between different objects, which is determined whether the bounding boxes of two objects overlap. If the bounding box of a_i has overlaps with the bounding box of b_j or c_z, turn to *Step 4*. Otherwise, turn to

TABLE I

SELECTED FEATURE MAPS IN SSD, FSSD AND OUR PROPOSED NETWORK WITH FEATURE FUSION MODULE. **AP** INDICATES AVERAGE POOLING, **FF** INDICATES FEATURE FUSION MODULE

SSD512	Conv4_3	FC7	Conv6_2	Conv7_2	Conv8_2	Conv9_2	Conv10_2
Resolution (px)	64×64	32×32	16×16	8×8	4×4	2×2	1×1

FSSD512	Concat_1	Concat_2	Concat_4	Concat_8	Concat_18	Concat_32	Concat_64
Resolution (px)	64×64	32×32	16×16	8×8	4×4	2×2	1×1

FS-SSD512	Final_1	Final_2	Final_3	Final_4	Final_5	Final_6	Ave pool
Resolution (px)	64×64	32×32	16×16	8×8	4×4	2×2	16×16

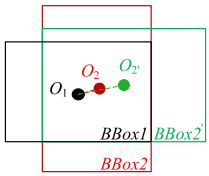


Fig. 6. The distance computing error caused by the overlap between bounding boxes.

Step 5. λ is a trade-off parameter between the FS-SSD model and the spatial context analysis method. Based on the above analysis, $C(a_i) \in [0.4, 0.6]$, $D(a_i, b_j)$, $D(a_i, c_z)$ as well as the final category confidence score $C'(a_i)$ is in the range of (0,1). To emphasize the influence of b_j on a_i and guaranteeing $C'(a_i)$ is not larger than 1 at the same time, theoretically, $\lambda \in [0.2, 0.4)$. Thus, the weight of spatial context analysis is no more than 0.4. Therefore, the influence weight of c_z on a_i is set to 0.4-λ. The exact value of λ will further be verified through Experiment IV in Section V.

Step 3: If there is no reliable object around the less reliable object a_i within a radius of d pixels. The more reliable objects beyond the radius of d to a_i, it is less likely for a_i to appear at current location. The final category confidence score C'' of a_i is decreased accordingly by:

$$C''(a_i) = C(a_i) - \lambda \times \left(\frac{e^{N(b_j)}}{e^{N(b_j)} + e^{N(c_z)}} \right) \tag{2}$$

where N counts the reliable objects whose distance to a_i is larger than d.

Step 4: Traditional distance measurement methods usually use Euclidean distance, which is calculated between the center points of two bounding boxes in object detection. However, Euclidean distance may cause calculation error when two bounding boxes have overlapped areas. As shown in Fig. 6, when there are existing bounding box overlaps, O_1 is the center of the bounding box of an object and O_2 and $O_{2'}$ are different bounding boxes of the other object. The distance D between two objects O_1 and O_2 is affected by the shape and the scale of the bounding boxes, and the further calculation of weighted distance will cause cumulative error. Enlightened by the distance computing in YOLOv2 [36], when two object bounding boxes have overlaps, we introduce intersection over union (IoU) to the distance measurement in spatial context analysis.

IoU is the ratio between the areas of overlap of two bounding boxes and their area of union, which is unrelated to the shape and size of the bounding box. IoU of the two bounding boxes A and B is computed as (3):

$$\text{IoU}(A, B) = \frac{S(A) \cap S(B)}{S(A) \cup S(B)} \tag{3}$$

where S is the area of the bounding boxes. The distance between object a_i and b_j is computed as follows:

$$D_{overlap}(a_i, b_j) = 1 - \text{IoU}\left(bbox\left(a_i\right), \quad bbox\left(b_j\right)\right) \tag{4}$$

Step 5: If the bounding boxes of a_i and b_j have no overlap, i.e., the IoU between them is 0, we still use Euclidean distance as the distance metric. D_{euc} is the Euclidean distance between two centers of the bounding boxes, and the final distance D is mapped to [0, 1] by min-max scaling:

$$D_{euc}(a_i, b_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{5}$$

$$D_{non\text{-}overlap}(a_i, b_j) = \frac{D_{euc}(a_i, b_j) - \min(D_{euc}(a_i, b_j))}{\max(D_{euc}(a_i, b_j)) - \min(D_{euc}(a_i, b_j))} \tag{6}$$

where (x_i, y_i) and (x_j, y_j) are the center coordinates of bounding boxes of a_i and b_j.

The objects in the following experiments contain three categories: pedestrian, biker and car. The widths and heights of these three kinds of objects range from 30 px to 80 px, and the instance number of pedestrians exceeds other two categories, so we take the size of pedestrian (50 px × 50 px) as a standard. Since the proposed spatial context analysis is performed based on the output image of the FS-SSD model, which has the same size as the original input image, the radius d is decided according to the pixel size of the objects and the sparsity among the objects in the input images. Statistically, the average distance among different object instances in the experimental dataset is 500 px. Thus, in the following experiment, the radius d is set to 500±50 px. The exact value of d will be determined through Experiment IV in Section V.

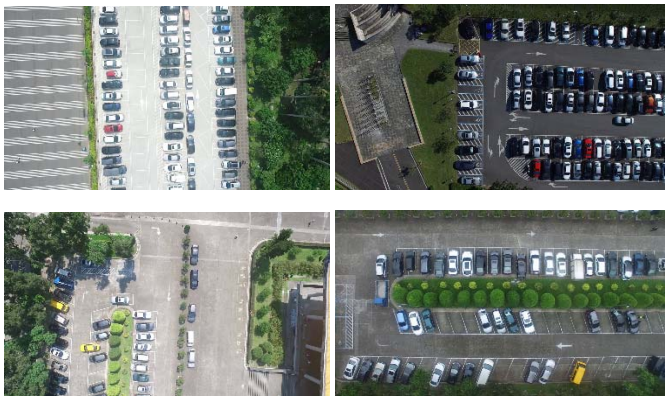


Fig. 7. Example scenes of CARPK dataset.

TABLE II
DATA DISTRIBUTION IN FOUR SCENES OF SDD

Scene	Images	Pedestrian	Biker	Car
bookstore	13053	1991	1017	118
hyang	23809	3358	1362	125
deathCircle	67947	2418	3598	560
little	18088	1622	2141	158

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Setup

1) Dataset Description: We have conducted the experiments on the following two representative and publicly available datasets of UAV images: the CARPK dataset [37] for car detection and the Stanford Drone Dataset (SDD) [38] for multiclass object detection. CARPK dataset is the first and the largest-scale drone view parking lot dataset that contains nearly 90,000 cars captured from four different parking lots. The resolution of the image is 1280 px $\times$ 720 px. Fig. 7 presents the examples of the CARPK dataset. We use 989 images for training and validation and 459 images for testing.

The Stanford Drone Dataset is the first large-scale dataset that has images and videos of various classes of objects that are moving and interacting on a real-world university campus. The whole dataset consists of six classes in eight unique scenes. However, since the ortho-image contains limited information for object detection, we use a subset of it with four scenes that captures many more objects, as shown in Fig. 8, named bookstore, hyang, deathCircle and little, respectively. In addition, the data distribution in the original dataset is seriously imbalanced. To balance the number of different objects, we divide the six classes into three groups, pedestrian, biker and car, according to their appearance and speed of movement.

The training and validation set contain 69673 images, 53224 images for testing. The number of objects in each category in each scene is reported in Table II, and the statistics of object size are shown in Fig. 9. The statistics show that the SDD dataset is very challenging due to the tiny size of three kinds of objects. All object instances have a size not larger than 0.2% of image size, and a considerable percentage of them are between 0.1-0.15%.

2) Evaluation Indicators: We use frames per second (FPS) to measure the detection speed, which represents the number

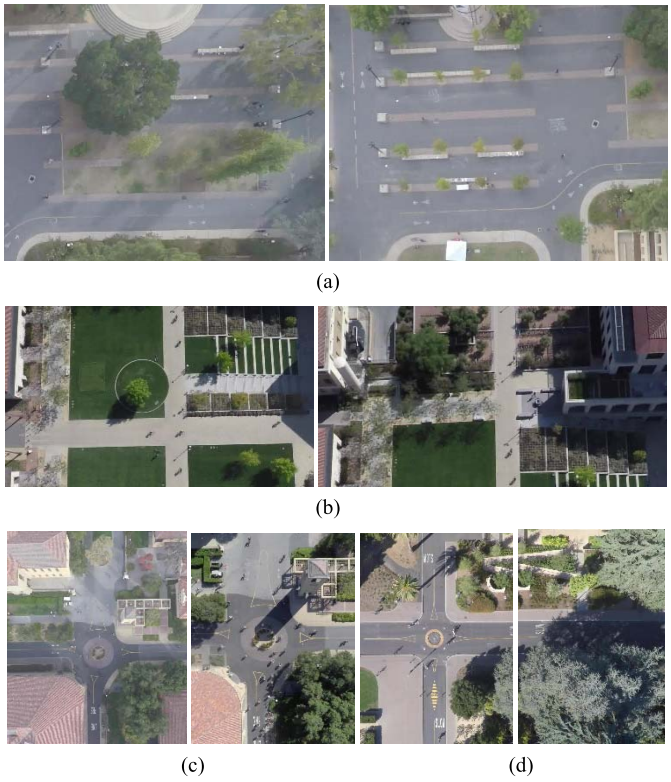


Fig. 8. Example scenes of Stanford Drone Dataset (SDD). (a) Example scenes of bookstore. (b) Example scenes of deathCircle. (b) Example scenes of hyang. (d) Example scenes of little.

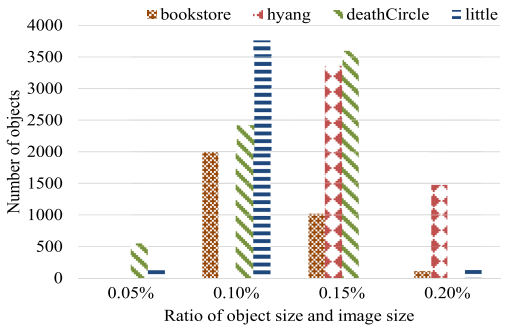


Fig. 9. Histogram of object size as a percentage of image area in four scenes of SDD.

of images the detection model can process per second with the specified hardware. In our experiments, we test the FPS of each method on a single GPU device. Mean average precision (mAP) is adopted as the criterion of detection accuracy, which is an indicator related to the IoU threshold. We take the most used threshold IoU = 0.5 in our experiments. In multiclass object detection, AP computes the area under the precision-recall curve and mAP is the average of the APs of multiple categories. Precision and recall can be defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

where *TP* means the true positives, i.e., IoU $>$ 0.5, *FP* indicates the false positives, and *FN* indicates the false negatives.

3) Implementation Details: The experimental platform is a PC with 3.40 GHz CPU, 16G RAM, Ubuntu 14.04 LTS

TABLE III
EFFECTS OF DIFFERENT DESIGN FACTORS ON THE SUBSET OF SDD. **AP** INDICATES AVERAGE POOLING, **FF** INDICATES FEATURE FUSION MODULE, **Deconv** INDICATES DECONVOLUTION MODULE

Methods	Dataset	AP	Feature Fusion	Deconv	mAP (%)
SSD512	SDD-subset	×	×	×	60.33
SSD512+FF	SDD-subset	×	Ele-sum	×	61.05
SSD512+FF	SDD-subset	×	Concat	×	61.74
SSD512+AP+FF	SDD-subset	√	Concat	×	63.26
SSD512+Conv6_2+FF	SDD-subset	×	Concat	×	62.02
SSD512+AP+Deconv	SDD-subset	√	×	√	63.52
SSD512+FF+Deconv	SDD-subset	×	Concat	√	63.83
FS-SSD512 (Ours)	SDD-subset	√	Concat	√	**65.84**

operating system. Our method is implemented by MATLAB 2014a and Python 2.7 based on Caffe [39] toolbox, accelerating by a NAVIDIA TITAN X (Pascal) GPU device with 12GB GPU memory, CUDA8.0 and cuDNN5.0.

Since the objects in PASCAL VOC and COCO datasets are much larger than the experimental datasets used in this paper, we change the aspect ratio according to the actual bounding boxes. Since most box ratios in CARPK and SDD datasets fall within a range of 1-2, we decide to use (1.5, 2.0) as the aspect ratio at every prediction layer. We apply NMS with a confidence threshold of 0.01, jaccard overlap of 0.45 per class and keep the top 200 detections per image.

We follow the same training policy as the original SSD. Data augmentation is done by randomly cropping the original image, and each sampled patch is [0.1, 1] of the original image size with aspect ratio between 0.5 and 2. The cropped patch is further horizontally flipped with probability of 0.5. We also use random photometric distortion, including hue and saturation shifts to simulate scenes with different illumination. All images are resized to $512 \times 512 \times 3$. For the training procedure, it has been proven that the best way of training a multiscale object detector is to make the training model and testing model have the similar size of inputs [40]. If a model trained on the large-scale object detection dataset, such as COCO, is used to detect the small objects directly, the problem of domain-shift is inevitable. Therefore, it is important to fine-tune on the small object datasets. Hence, we take the original SSD model trained on PASCAL VOC 2007, PASCAL VOC 2012 and COCO datasets as a pretrained model. Then, we fine-tune the model on the CARPK dataset and a subset of SDD dataset, respectively. The learning rate is set to 10^{-4} and decrease 10% with step size of 10^4, the momentum parameter is chosen as 0.9, and the weight decay is 0.0005. The max iteration is set to 80000.

The training objective is to minimize the weighted sum of Smooth L1 loss [16] L_{loc} for localization and Softmax loss L_{conf} for classification confidence in (9), which is the same as the loss function in SSD [20].

$$L(x, c, p, g) = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, p, g)) \quad (9)$$

where N is the number of matched default boxes for random object x, c is the class confidence score, p is the predicted bounding box, and g is the ground-truth bounding box. The hyperparameter α is set to 1 by cross validation.

B. Experiment I: Ablation Study of the Proposed FS-SSD Network on the Subset of SDD

To understand the effectiveness of the different network module, we run models with different settings on a subset of SDD and record their evaluations in Table III.

First, we test the effect of the feature fusion module. Table III shows that the pure SSD512 model with different ways of feature fusion can get different results. Using concatenation to fuse features can get 61.74% mAP (row 3) while element-wise summation can only achieve 61.05% (row 2). In the following experiments, we use concatenation to fuse features for better performance. Then, we add the additional Conv6_2 with average pooling layer to the SSD512 model together with feature fusion module. The mAP has an improvement of 1.52% (row 4). To further verify the effectiveness of the use of average pooling, we remove the average pooling layer, only keep the additional Conv6_2 to SSD512 with feature fusion module (row 5). However, the mAP has an improvement of only 0.28% compared to the combination of SSD512 and future fusion module, and has a decline with 1.24% compared to the combination of SSD512, average pooling and future fusion module. The results demonstrate that the use of additional convolutional layer with average pooling contributes more to the detection accuracy than additional convolution itself. Next, we add the average pooling operations to the deconvolution module. Compared with the combination of average pooling and the feature fusion module, the combination of average pooling and the deconvolution module results in a higher mAP with 63.52% (row 6). After that, we replace the average pooling with the deconvolution module, and this time mAP achieves 63.83% (row 7), which is slightly higher than the previous models. Finally, we propose the feature fusion and scaling-based SSD that involve both the feature fusion module and the deconvolution module with average pooling. Our model gets the highest mAP with 65.84% (row 8).

C. Experiment II: Comparison between the Proposed FS-SSD and the State-of-the-Art on PASCAL VOC 2007 Benchmark

In this experiment, we compare the proposed FS-SSD512 model with the baseline method and other state-of-the-art detectors on the PASCAL VOC 2007 benchmark. Since our methods are mainly modified based on FSSD512 [31], FSSD512 is chosen as the baseline method in the following experiments. Other deep learning-based

TABLE IV

COMPARISON OF THE PROPOSED FS-SSD WITH STATE-OF-THE-ART DETECTORS ON THE PASCAL VOC 2007 TEST. ∗ INDICATES BASELINE METHOD

Methods	Data	mAP (%)	FPS	#Boxes
Faster R-CNN(VGG-16)	07+12	73.20	7.60	300
R-FCN(ResNet-101)	07+12	79.50	10.40	300
DSSD513(ResNet-101)	07+12	81.50	5.50	43688
FSSD512*(VGG-16)	07+12	80.90	23.74	24564
SSD512(VGG-16)	07+12	79.80	18.26	24564
SSD300(VGG-16)	07+12	77.20	35.25	8732
YOLOv3(DarkNet-53)	07+12	79.61	46.94	10647
FS-SSD512(VGG-16)	07+12	81.30	18.29	24528

TABLE V

COMPARISON OF THE PROPOSED FS-SSD WITH STATE-OF-THE-ART DETECTORS ON THE CARPK DATASET, ∗ INDICATES BASELINE METHOD

Methods	mAP (%)	FPS	#Boxes
Faster R-CNN(VGG-16)	84.80	7.60	300
R-FCN(ResNet-101)	86.13	10.40	300
DSSD513(ResNet-101)	87.27	5.50	43688
FSSD512*(VGG-16)	87.59	23.74	24564
SSD512(VGG-16)	86.82	18.26	24564
SSD300(VGG-16)	82.72	35.25	8732
YOLOv3(DarkNet-53)	86.01	46.94	10647
FS-SSD512(VGG-16)	89.52	18.29	24528

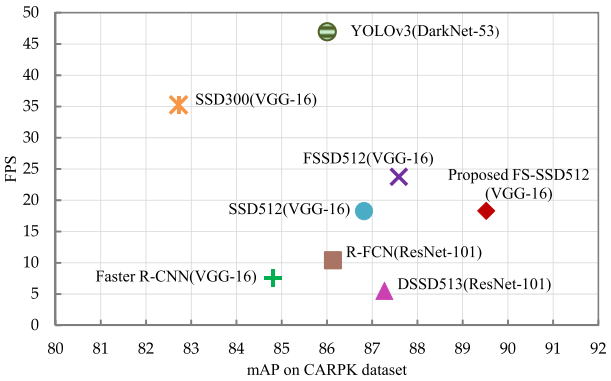


Fig. 10. Speed and accuracy comparison of the proposed method with the state-of-the-art methods on the CARPK dataset. Our FS-SSD512 model is more accurate compared with other models with a little compromise on speed.

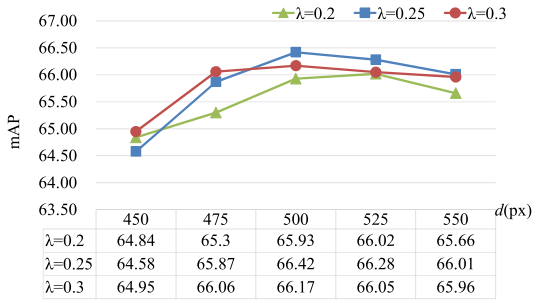
Fig. 11. Comparison between the original SSD512 and the proposed FS-SSD512 on the CARPK dataset. Bounding boxes with scores of 0.5 or higher are drawn. (a) Results of the original SSD. (b) Results of our method.

detectors include two-stage detectors Faster R-CNN [17] and R-FCN [18], one-stage detectors SSD [20], DSSD513 [28] and YOLOv3 [41]. All detectors are pretrained on PASCAL VOC 2007 and the 2012 dataset. Table IV shows that the proposed FS-SSD512 model achieves the second highest mAP with 81.32% among seven deep models, which is only 0.2% lower than the best performance of DSSD513, while the detection speed of FS-SSD512 is nearly 2.5 times faster than DSSD513. The proposed FS-SSD512 model exceeds FSSD512 in accuracy with a little drop in FPS. The powerful backbone network ResNet-101 [42] as well as the deconvolution module contribute to the good results of DSSD513. YOLOv3 is the latest version of YOLO. Although the detection speed is quickest among the six networks, the accuracy on small objects is far from satisfactory. The SSDs results cited here are the version updated by the author after paper publication.

D. Experiment III: Comparison Between the Proposed FS-SSD and the State-of-the-Art on CARPK Dataset

In this experiment, we verify the effectiveness of the proposed FS-SSD on the CARPK dataset. Table V shows that our FS-SSD512 with the VGG-16 backbone achieves 89.52% mAP, which improves 1.93 points compared with the second-best network FSSD512 with a little compromise on the detection speed. However, the detection speed of 18.29 FPS is acceptable for real-time detection.

The speed of the proposed detector in comparison to state-of-the-art detectors is shown in Fig. 10. The deeper architecture (ResNet) used in R-FCN and DSSD513 makes the model



	450	475	500	525	550
λ=0.2	64.84	65.3	65.93	66.02	65.66
λ=0.25	64.58	65.87	66.42	66.28	66.01
λ=0.3	64.95	66.06	66.17	66.05	65.96

Fig. 12. The detection results under different parameter settings on the subset of SDD.

TABLE VI
COMPARISON OF SPEED AND ACCURACY OF STATE-OF-THE-ART DETECTORS ON THE SUBSET OF SDD DATASET. **SCA** INDICATES SPATIAL CONTEXT ANALYSIS, ∗ INDICATES BASELINE METHOD

Methods	mAP (%)	FPS	#Boxes
Faster R-CNN(VGG-16)	59.63	7.60	300
R-FCN(ResNet-101)	61.30	10.40	300
DSSD513(ResNet-101)	63.08	5.50	43688
FSSD512*(VGG-16)	64.19	23.74	24564
SSD512(VGG-16)	60.33	18.26	24564
SSD300(VGG-16)	58.24	35.25	8732
YOLOv3(DarkNet-53)	62.23	46.94	10647
FS-SSD512(VGG-16)	**65.84**	18.29	24528
FS-SSD512+SCA(VGG-16)	**66.42**	18.25	24528

more accurate but increases the computational complexity. Fig. 11 presents the comparison between the original SSD and the proposed FS-SSD on the CARPK dataset. Fig. 11(a) is the detection result of the original SSD, which misses many of the object instances with dense attribution, shadow or occlusion. The false positive rate is quite decreased with the proposed FS-SSD in Fig. 11(b).

E. Experiment IV: Detection Results Under Different Parameter Settings in Spatial Context Analysis for Object Redetection

In this experiment, we try to find the best parameter setting in a spatial context analysis for object redetection, mainly including the trade-off parameter λ between the detection model and the spatial context analysis method and the distance threshold d between different object instances. According to the analysis in Section IV, the trade-off parameter λ is in the range of 0.2 to 0.4, and the average distance between different object instances is approximately 500 px. Thus, λ is set to 0.2, 0.25 and 0.3 and d is set to the range of 450 px to 550 px at intervals of 50 px. As Fig. 12 shows, the results indicate that our method is sensitive to the choice of λ and d on the subset of SDD. Basically, the mAP increases when d changes from 450 px to 500 px, and the best performance with mAP of 66.42% is achieved when the λ is set to 0.25, and d is set to 500 pixels. When $\lambda = 0.2$, the detection accuracy is generally below the other two group results, regardless of the choice of d. When $\lambda = 0.3$, although the performance is superior to the other two groups when d is 450 px and 475 px, the accuracy decline keeps increasing the value of d. The best results will be compared with the state-of-the-art detectors in the next experiment.

(a) (b)

Fig. 13. FS-SSD512 vs. FS-SSD512 with spatial context analysis. Both models are trained with VOC 07 + 12 and MS COCO datasets. Bounding boxes with a score of 0.5 or higher are drawn. (a) The detection results of FS-SSD512. (b) The corresponding detection results of FS-SSD512 with spatial context analysis.

TABLE VII

DETECTION RESULTS OF STATE-OF-THE-ART DETECTORS ON THE SUBSET OF SDD DATASET. ALL DETECTORS ARE PRE-TRAINED ON PASCAL VOC 2007, 2012 AND MS COCO DATASET. **SCA** INDICATES SPATIAL CONTEXT ANALYSIS

Methods	Pretrained Datasets	Backbone	Input	Car	Pedestrian	Biker	mAP (%)
Faster R-CNN	07+12+COCO	VGG-16	~1000×600	58.57	67.08	53.24	59.63
R-FCN	07+12+COCO	ResNet-101	~1000×600	61.57	67.42	54.92	61.30
DSSD513	07+12+COCO	ResNet-101	513×513	62.75	71.69	54.80	63.08
FSSD512(baseline)	07+12+COCO	VGG-16	512×512	65.44	71.98	55.14	64.19
SSD512	07+12+COCO	VGG-16	512×512	60.55	67.54	52.89	60.33
SSD300	07+12+COCO	VGG-16	300×300	57.26	66.58	50.89	58.24
YOLOv3	07+12+COCO	DarkNet-53	416×416	61.83	73.52	51.33	62.23
FS-SSD512	07+12+COCO	VGG-16	512×512	66.49	73.08	57.95	65.84
FS-SSD512+SCA	07+12+COCO	VGG-16	512×512	**66.72**	**74.10**	**58.44**	**66.42**

F. Experiment V: Comparison Between the Proposed FS-SSD With Spatial Context Analysis and the State-of-the-Art on the Subset of SDD Dataset

In this experiment, we compare the objective detection results of the proposed FS-SSD with spatial context analysis with six state-of-the-art detectors on the subset of SDD. As Table VI shows, our FS-SSD model with 512 × 512 input has achieved 65.84% mAP and is superior to the latest FSSD512 by 1.65 points. By adding the spatial context analysis for redetection, the accuracy of the proposed model increases from 65.84% to 66.42%, and it outperforms other deep networks. For the detection speed, our FS-SSD512 model maintains a relative speed the same as the original SSD512, which basically meets the requirements of real-time detection. Table VII is the detailed detection results of different classes in the SDD dataset. Three classes achieve the best accuracy on our feature fusion and scaling-based SSD with spatial context analysis. The results demonstrate that in multiclass small object detection, the interplay between different object instances helps to improve the performance.

G. Experiment VI: Subjective Comparison of Detection Results Before and After Using Spatial Context Analysis

To demonstrate the effectiveness of our feature fusion and scaling-based SSD with the spatial context analysis method, we present the detection results before and after using spatial context analysis. As shown in Fig. 13, the left side is the proposed FS-SSD512 model, and the right side is the FS-SSD512 with spatial context analysis. After considering the influence around a specific object, the false negative rate is decreased since there are more objects can be correctly detected. For example, in the first row, there are only one biker and nine pedestrians detected by the FS-SSD512 model, but we can clearly see that the other biker is just located on the top left of the detected biker, and there are obviously more pedestrians in this image. However, considering the influence of the surrounding reliable objects, the other biker and more pedestrians around these reliable objects have been detected, i.e., the less reliable objects achieve an improvement with different degrees in their confidence scores.

VI. DISCUSSION

In this paper, we propose a feature fusion and scaling-based single shot detector (FS-SSD) with spatial context analysis to tackle the problem of small object detection in UAV images. On the premise of real-time detection, we improve the detection accuracy compared with other deep learning-based detectors. Based on FSSD [31], we add an extra scaling branch of the deconvolution module with average pooling operation and redesign the concatenation of multilevel feature maps. The entire network is deeper but narrower so that the representative ability has been strengthened by joint prediction of two feature pyramids. Aside from the proposed FS-SSD, we take the spatial relationships of multiclass objects into consideration to further improve the detection accuracy. The interclass and intraclass distances between different object instances are computed as spatial context to exert influence on the final detection results.

Six experiments are conducted to verify the effectiveness of our method. In Experiment I, an ablation study is performed to analyze different combinations of network modules. As shown in Table III, our FS-SSD network achieves the highest mAP with 65.84%, indicating that the extra branch of the deconvolution module with average pooling together with feature fusion can fully exploit the features of small objects. In Experiment II, the results in Table IV show that the proposed FS-SSD512 model surpasses the baseline method FSSD512, achieving the second highest mAP on the PASCAL VOC 2007 benchmark among seven state-of-the-art deep detectors with a little drop in FPS. In Experiment III, we compare the proposed FS-SSD with six state-of-art methods on the CARPK dataset, and our model achieves the highest mAP with comparable detection speed. By absorbing the advantage of design concept in state-of-the-art models, the proposed FS-SSD achieves higher detection accuracy due to the improvement of the network representative ability. Fig. 11 is the subjective comparison between our FS-SSD and the original SSD [20]. Our FS-SSD can effectively decrease the false positive rate of densely distributed objects in different scenes. In Experiment IV, the best parameter setting in spatial context analysis is explored on the subset of SDD. Fig. 12 shows that best mAP with 66.42% is achieved when trade-off parameter λ between the FS-SSD model and the spatial context analysis is set to 0.25, and the distance threshold d between different object instances is set to 500 pixels. We further compare this result with other state-of-the-art methods in Experiment V. In general, after incorporating the spatial context analysis into the FS-SSD model, the performance improves compared with the results without spatial context analysis and other

six state-of-the-art methods in Table VI and VII. Table VI shows that our FS-SSD model without spatial context analysis can also outperform FSSD512 in mAP with a promotion of 1.65 points, which indicates that our network has better representative ability for small objects. The subjective detection results are shown in Fig. 13, which further proves the effectiveness of spatial context analysis for multiclass small object detection.

VII. Conclusions and Future Work

With the rapid development and wide application of UAV photography, how to accurately and quickly detect the small objects in UAV images is a challenging problem in the research area of computer vision. In this paper, we aim to improve the network architecture based on FSSD, an improvement of the original SSD. We propose a method for small object detection in UAV images by using a feature fusion and scaling-based single shot detector with spatial context analysis to improve the detection accuracy. First, an ablation study with different settings is conducted on the pure SSD network to verify the effectiveness of the improved feature fusion module and the extra branch of the deconvolution module with average pooling. From the results, we can see that the proposed FS-SSD model achieves the highest mean average precision rate with the advantage of full utilization of the feature neighborhood information and multiscale feature maps. Then, we compare the detection results of FSSD on the PASCAL VOC benchmark to prove the effectiveness of our model. Next, we compare the proposed FS-SSD with six state-of-the-art methods on the CARPK dataset. The results demonstrate that the proposed FS-SSD can get superior detection accuracy due to the absorption of essence from the state-of-the-arts to improve the network representative ability. Since it is not enough to classify a small object relying only on its features learned by deep learning detectors, we finally involve the spatial context analysis in object redetection by taking the interplay of multiclass objects within a certain distance into consideration. The pixel distances between intraclass and interclass object instances are computed as the spatial context to redetect the object at a fine level. We investigate the parameter settings in spatial context analysis, including the trade-off parameter λ between the detection model and the spatial context analysis method and the distance threshold d between different object instances. The experimental results indicate that the proposed FS-SSD outperformed the six state-of-the-art methods in accuracy on the subset of SDD with speed comparable to other detectors. The subjective comparison results on the subset of the SDD dataset further prove that based on the proposed FS-SSD, spatial context analysis can greatly contribute to the detection accuracy when λ and d are set to proper values. In our future work, a more effective and accurate object detection architecture will be considered for UAV imagery object detection.

In the future, we plan to replace the backbone network with a much stronger network, such as ResNet [42], to obtain better feature representation with the advantage of computational simplicity. Meanwhile, we will apply videos rather than still images, which raises a higher demand on real time detection. In addition, a better loss function, such as distillation loss [43], and the choice of training data are also worth exploring.

References

[1] K. Dimitropoulos, P. Barmpoutis, and N. Grammalidis, "Spatio-temporal flame modeling and dynamic texture analysis for automatic video-based fire detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 2, pp. 339–351, Feb. 2015.

[2] M. B. Bejiga, A. Zeggada, A. Nouffidj, and F. Melgani, "A convolutional neural network approach for assisting avalanche search and rescue operations with UAV imagery," *Remote Sens.*, vol. 9, no. 2, p. 100, 2017.

[3] X. Wang, A. Chowdhery, and M. Chiang, "Networked drone cameras for sports streaming," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst.*, Atlanta, GA, USA, Jul. 2017, pp. 308–318.

[4] I. Colomina and P. Molina, "Unmanned aerial systems for photogrammetry and remote sensing: A review," *ISPRS J. Photogram. Remote Sens.*, vol. 92, pp. 79–97, Jun. 2014.

[5] L. Mauelshagen, "Low altitude aerial photography," *Photogramm. Rec.*, vol. 12, no. 68, pp. 239–241, 1986.

[6] S. Kamate and N. Yilmazer, "Application of object detection and tracking techniques for unmanned aerial vehicles," *Procedia Comput. Sci.*, vol. 61, pp. 436–441, Nov. 2015.

[7] L. Wang, F. Chen, and H. Yin, "Detecting and tracking vehicles in traffic by unmanned aerial vehicles," *Automat. Construct.*, vol. 72, pp. 294–308, Dec. 2016.

[8] T. Moranduzzo and F. Melgani, "Detecting cars in UAV images with a catalog-based approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6356–6367, Oct. 2014.

[9] X. Wen, L. Shao, W. Fang, and Y. Xue, "Efficient feature selection and classification for vehicle detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 508–517, Mar. 2015.

[10] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[11] M. Radovic, O. Adarkwa, and Q. Wang, "Object recognition in aerial images using convolutional neural networks," *J. Imag.*, vol. 3, no. 2, pp. 21–29, 2017.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, and S. Ozair, "Generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014, pp. 2672–2680.

[13] X. Shu, G.-J. Qi, J. Tang, and J. Wang, "Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *Proc. ACM Int. Conf. Multimedia*, Brisbane, QLD, Australia, Oct. 2015, pp. 35–44.

[14] J. Tang, X. Shu, Z. Li, G. J. Qi, and J. Wang, "Generalized deep transfer networks for knowledge propagation in heterogeneous domains," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 12, no. 4s, pp. 68:1–68:22, 2016.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, Columbus, OH, USA, Jun. 2014, pp. 580–587.

[16] R. Girshick, "Fast R-CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, Boston, MA, USA, Jun. 2015, pp. 1440–1448.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2015, pp. 91–99.

[18] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 379–387.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788.

[20] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.

[21] J. Huang *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 7310–7311.

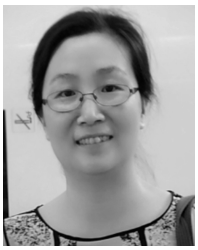
[22] M. Bar, "Visual objects in context," *Nature Rev. Neuro-Sci.*, vol. 5, no. 8, pp. 617–629, 2004.

[23] P. Carbonetto, N. de Freitas, and K. Barnard, "A statistical model for general contextual object recognition," in *Proc. Eur. Conf. Comput. Vis.*, Prague, Czech Republic, May 2004, pp. 350–362.

[24] G.-J. Qi, "Hierarchically gated deep networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 2267–2275.
[25] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.
[26] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Zürich, Switzerland, Sep. 2014, pp. 740–755.
[27] M. Lin, Q. Chen, and S. Yan. (2013). "Network in network." [Online]. Available: https://arxiv.org/abs/1312.4400
[28] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. (2017). "DSSD: Deconvolutional single shot detector." [Online]. Available: https://arxiv.org/abs/1701.06659
[29] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun. (2018). "DetNet: A backbone network for object detection." [Online]. Available: https://arxiv.org/abs/1804.06215
[30] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 2117–2125.
[31] Z. Li and F. Zhou. (2017). "FSSD: Feature fusion single shot multibox detector." [Online]. Available: https://arxiv.org/abs/1712.00960
[32] Y. L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 2559–2566.
[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, Jul. 2015, pp. 448–456.
[34] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1–9.
[35] J. Leng and Y. Liu, "An enhanced SSD with feature fusion and visual reasoning for object detection," *Neural Comput. Appl.*, vol. 2, pp. 1–10, Apr. 2018.
[36] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, Honolulu, HI, USA, Jul. 2017, pp. 6517–6525.
[37] M.-R. Hsieh, Y.-L. Lin, and W.-H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 4165–4173.
[38] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 549–565.
[39] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, Orlando, FL, USA, Nov. 2014, pp. 675–678.
[40] H. Qin, X. Li, Y. Wang, Y. Zhang, and Q. Dai, "Depth estimation by parameter transfer with a lightweight model for single still images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 748–759, Apr. 2017.
[41] J. Redmon and A. Farhadi. (2018). "YOLOv3: An incremental improvement." [Online]. Available: https://arxiv.org/abs/1804.02767
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[43] R. Mehta and C. Ozturk. (2018). "Object detection at 200 frames per second." [Online]. Available: https://arxiv.org/abs/1805.06361

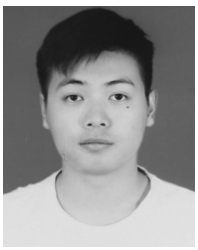
Xi Liang (M'18) is currently pursuing the master's degree with the Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology.

Her research interests include image recognition and object detection for transportation.



Jing Zhang (M'14) received the Ph.D. degree in circuit and system from the Beijing University of Technology in 2008. She is currently a Professor with the Faculty of Information Technology, Beijing University of Technology, where she is also with the Beijing Key Laboratory of Computational Intelligence and Intelligent System. She is also a Research Scholar with the Department of Computer Science, The University of Texas at San Antonio (UTSA). She has authored more than 60 journal papers and has written four book chapters. Her current research interests include image processing, image recognition, and image retrieval.

Li Zhuo received the Ph.D. degree in pattern recognition and intelligent system from the Beijing University of Technology in 2004. She is currently a Professor with the Faculty of Information Technology, Beijing University of Technology, where she is also with the Beijing Key Laboratory of Computational Intelligence and Intelligent System and with the Collaborative Innovation Center of Electric Vehicles in Beijing. She published over 180 refereed journals and conference papers, and has written six book chapters. Her current research interests include image/video processing, image recognition, and medical image processing.



Yuzhao Li is currently pursuing the master's degree with the Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology.

His research interest includes image recognition for transportation.

Qi Tian (S'95–M'96–SM'03–F'16) received the B.E. degree in electronic engineering from Tsinghua University in 1992, the M.S. degree in ECE from Drexel University in 1996, and the Ph.D. degree in ECE from the University of Illinois at Urbana-Champaign (UIUC) in 2002.

He was a tenure-track Assistant Professor from 2002 to 2008 and a tenured Associate Professor from 2008 to 2012. From 2008 to 2009, he took one-year Faculty Leave at Microsoft Research Asia (MSRA) as a Lead Researcher with the Media Computing Group. He is currently a Full Professor with the Department of Computer Science, The University of Texas at San Antonio (UTSA), San Antonio, TX, USA, and the Chief Scientist of Computer Vision at the Huawei Technologies Noah's Ark Laboratory. He has published over 360 refereed journals and conference papers. His research interests include multimedia information retrieval, computer vision, and pattern recognition.

Dr. Tian was a Co-Author of the Best Paper in ACM ICMR 2015, the Best Paper in PCM 2013, the Best Paper in MMM 2013, the Best Paper in ACM ICIMCS 2012, the Top 10% Paper Award in MMSP 2011, the Best Student Paper in ICASSP 2006, the Best Student Paper Candidate in ICME 2015, and the Best Paper Candidate in PCM 2007. He received the 2017 UTSA President's Distinguished Award for Research Achievement, the 2016 UTSA Innovation Award, the 2014 Research Achievement Awards from the College of Science, UTSA, the 2010 Google Faculty Award, and the 2010 ACM Service Award. He is an Associate Editor of many journals and is in the Editorial Board of the *Journal of Multimedia* (JMM) and the *Journal of Machine Vision and Applications* (MVA).